ESD-TDR- 64-96

TM-3869

FINAL REPORT ON THE ROUT DOCUMENT RETRIEVAL SYSTEM

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR- 64-96

MAY 1964

J. F. Rial

Prepared for

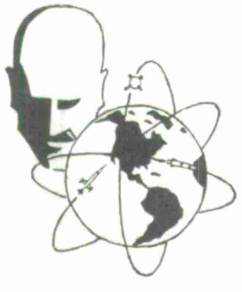
438L (ADVANCED INTELLIGENCE DATA SYSTEM)

ELECTRONIC SYSTEMS DIVISION

AIR FORCE SYSTEMS COMMAND

UNITED STATES AIR FORCE

L. G. Hanscom Field, Bedford, Massachusetts



Project 438L

Prepared by

THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-2390

FINAL REPORT ON THE ROUT DOCUMENT RETRIEVAL SYSTEM

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR- 64-96

MAY 1964

J. F. Rial

Prepared for

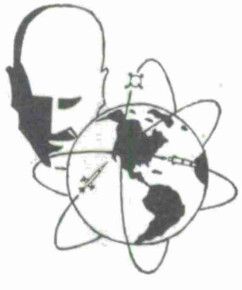
438L (ADVANCED INTELLIGENCE DATA SYSTEM)

ELECTRONIC SYSTEMS DIVISION

AIR FORCE SYSTEMS COMMAND

UNITED STATES AIR FORCE

L. G. Hanscom Field, Bedford, Massachusetts



Project 438L

Prepared by

THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF19(628)-2390

FOREWORD

The evaluation of the ROUT program was performed by four people. René Radner assisted in the rating of messages and in the preparation of queries. Christianne Schmidt, a summer student, prepared several of the histograms, did the curve-fitting in Section VII, and performed the calculations needed to construct Figs. 16 and 17. Dianne Budde and the author shared the rest of the work. The design of the evaluation and the plan of the final report are the work of the author. The ROUT program was written by Terence Dewey, Hugh Lynch, William Aldrich, René Radner, Richard Justice of Department D-17, Paul Collard on loan from Systems Development Corporation (SDC), Richard Baust of Department D-19, Eugene Famolari of Department D-17, Israel Feldman of Department D-27, and Charles Krisher, a summer student, who prepared the Metric Search routines. Dr. Stanley Frank of Department D-17 gave considerable help in formulating the statistical tests.

ABSTRACT

The evaluation of the ROUT (Retrieval of Unformatted Text) document retrieval system is a detailed examination of various retrieval errors in the context of different query-processing methods. Many of the methods used to extend the power of the basic coordinate indexing mechanism are compared. Three different evaluation techniques are used in the study. Based on the analysis, conclusions are drawn about the relative merits of several query-processing methods, and some general comments are made about the limitations of coordinate indexing as applied to the document retrieval problem.

SECTION I

INTRODUCTION

ROUT (Retrieval Of Unformatted Text) is a document retrieval system based upon coordinate indexing. Throughout this paper, we shall use the term coordinate indexing in the following sense: suppose we have a set D of documents, each of which is indexed by one or more items from a set I, which may contain words, phrases, dates, etc. Suppose each item is then associated with the documents containing that item. The pair (D, I) is a coordinate indexing system if retrieval is carried out by the Boolean combination of document sets as prescribed by the corresponding Boolean combination of the items associated with the document sets. The primary inadequacies of coordinate indexing have been pointed out by many critics, Bar Hillel chief among them. However, the attacks leveled at this type of document retrieval have not, to the author's knowledge, been supported by thorough, quantitative analyses of actual retrieval systems. There are, on the other hand, strong advocates of coordinate indexing, and from this quarter, too, there has been a discouraging lack of evidence to substantiate their claims. We will not engage in this battle. The purpose of this paper is to present facts about the performance of a particular coordinate indexing system. The reader is left to draw his own conclusions about the merits and inadequacies of coordinate indexing in general.

There are many ways of evaluating document retrieval systems. The ROUT evaluation is founded on two techniques: one is designed to rate the system against an artificial retrieval method whose characteristics are precisely determined, the other is designed primarily as a means for comparing different query-processing methods and conducting detailed analyses of the system errors.

The latter technique was performed in three ways: one based on a categorization of the document file, another on sampling, and the third on rating every message in the file against every question used.

Various statistical tests were applied to the evaluation data and results are chiefly stated in terms of the rejection of null hypotheses at 5 and 10 percent significance levels. Tables and graphs are included to exhibit general features of system performance.

SECTION II

DESCRIPTION OF ROUT PROCESSING CAPABILITIES

The ROUT system is programmed for the IBM 1410 with a 1301 disc file. The document file consists of NORAD intelligence messages automatically indexed (on an IBM 7090) against a fixed dictionary of key words and key phrases. The dictionary is organized into a thesaurus which groups key words by synonym and subordinate relationships. A key word may have synonyms and subordinates, synonyms and no subordinates, subordinates and no synonyms, or neither synonyms nor subordinates. The thesaurus gives rise to three query processing methods. (A query is any Boolean combination of keywords or key phrases. A key word in a given query will, for convenience, be called a query word.) The query may be processed without making use of either synonyms or subordinates of the query words. Such a query will be called Unexpanded, and the associated query process designated No Expansion. The query may be expanded by logically adding to each query word its synonyms, thus giving rise to the process designated Synonym Expansion. Finally, a query may be expanded by logically adding to each query word both its synonyms and subordinates. The associated query process is called Complete Expansion.

A query is entered into the system at the 1410 console typewriter. The program can be directed to process the query in any of the forms mentioned above. If the query is to be expanded, then the synonyms, or synonyms and subordinates, of each query word will be printed out at the console typewriter. The querist may now modify his query by deleting any or all of the synonyms and subordinates of the query words. This process, called Delete, gives to the querist a certain semantic freedom. The querist may now order a search, and

the program will generate, for each synonym and subordinate group in the query, a list of the numbers of those messages containing any of the key words in the group. The resulting lists are now multiplied or added together exactly as the key words from which they are derived are multiplied or added together in the query. The querist is presented with the number of messages generated by the list processing. He can then request either a printout of the full texts of the messages, or a printout, for each message, of the key words with which the message is indexed. The querist, in the latter case, selects messages on the basis of these abstracts. The process is called Bibliography, or merely Bib.

A document association technique can be used to extend any of the expansion processes described previously. The distance d between two documents A_1 and A_2 is defined as:

$$d(A_1, A_2) = 1 - \frac{\text{Number of key words } A_1 \text{ and } A_2 \text{ have in common}}{\text{Number of key words in either } A_1 \text{ or } A_2}$$

There are a number of ways of using this function in retrieval; only one was chosen for evaluation in the ROUT study. A query is treated by some combination of the previously described processes, yielding a set of messages. The five closest messages to each of these retrieved messages are obtained. The process is called Metric Search. It can be repeated on selected messages retrieved by an extension of the Metric Search process, and the resulting process is designated Metric Search-Metric Search (Metric Search^2).

The six query-processing methods defined above can be used in a large variety of combinations. Any mixture of the Delete, Bibliography, Metric Search, and Metric Search^2 processes can be used with any of the three expansion processes. ROUT can thus process any given query in $3(2^4) = 48$ distinct ways.

It was not necessary, however, to evaluate all of these combinations in order to decide which combinations provide the greatest capability. The selection of the combinations chosen for study will be defended in Section V. For the present, we observe that ROUT contains a significant cross section of the query-processing techniques available in coordinate indexing systems.

SECTION III

DESCRIPTION OF EVALUATION TECHNIQUES

The evaluation of any document retrieval system presents many basic difficulties. The document file is, of course, a more or less well-defined entity. The collection of subsets of the document file constitutes a fixed, finite answer space. This space can be sampled, and it can often be rigidly categorized for the purposes of more efficient retrieval. The question space, on the other hand, cannot be defined even generally, and thus can be neither categorized nor sampled in a precise manner. One of the challenges in designing a retrieval experiment is to choose a class of questions which, in some way, represents the kind of question that will typically be put to the system. The severity of this problem is eased somewhat if the customer picks the questions. In the case of ROUT the problem was instead aggravated because there was no customer, and, hence, no criterion for choosing representative questions. It was thus decided to ask all kinds of questions: general and specific, qualitative and quantitative, and about localities, personalities, political and the military situation. The work of making up 90 questions was divided among three people, who rated the messages as relevant or irrelevant to the questions. Each message was rated only once with respect to a given question, and the three people often rated messages with respect to their own questions.* The effects of these divisions of labor on the evaluation data are given in Section VI.

*Each person rating messages first read the material in the categories they handled.

To study the effects of query reformulation, each of the questions was translated into two queries, or Boolean combinations of key words. (The Delete process is, in a sense, also a query reformulation process.) Each question was translated by the person originating the question.

The queries were treated by different combinations of the processes described in Section II, yielding, for a given query and a given combination, a set of messages we shall call a computed reply. Suppose, for some question, we have a computed reply containing N messages, and suppose T of those messages were previously judged to be relevant to the question. An evaluation technique can be based upon the following consideration: if N messages are selected at random from the message file, then what is the probability that at least T of those messages are relevant to the same question? The random sampling of a document file defines an elementary type of retrieval system with very minimal efficiency. It is, nevertheless, a system whose capabilities are completely and rigorously determined, and thus can be used as an absolute standard of comparison. The question asked in the ROUT evaluation was: How does ROUT's ability to retrieve relevant messages compare with a system which retrieves documents by random sampling? Different query processes were examined in the light of this question.

Some of the performance measures described in Section IV were studied under three different conditions. In one case, the evaluation data were derived from querying three samples of 100 messages each. One of the samples was drawn strictly at random. The other two were chosen by picking, at random, some message in the file, and then taking that and the following 99 messages. The purpose of sampling in different ways was to discover if results were influenced by the uniqueness of the file structure. Another condition under which data were collected was derived from spliting up the message file into eight mutually exclusive sets, the sum of these sets exhausting the file. Each message

set, or category, deals with a particular area of the world. The questions used throughout the evaluation were originally made up with respect to these categories. Each question dealt with one and only one category, and the rating of messages on the basis of relevancy to the question was founded on the assumption that every message not contained in the given category was irrelevant to the question. The effect of this was to create eight mutually independent samples which could, for the purposes of analysis, be combined in many ways. Many of the evaluation results are based on a study of how system performance varies over different category combinations. The final condition under which data were collected depended upon rating each message in the modified message file against a selected subset of the queries. The results of the three different kinds of testing were compared and, based on the comparison, a judgment was made on how best to evaluate a retrieval system operating on a larger document file (see Sections VII and VIII).

During the evaluation, the message file underwent a change which will now be described. About 11 percent of the original message file consisted of multitopic, or composite, messages. Each of these messages was coded relative to the categories it dealt with. The presence of composite messages in the document file of a coordinate indexing system gives rise to a unique phenomenon which was separately studied, namely, that an irrelevant composite message can be retrieved because the words (or phrases) in a query are distributed among the different topics of the message. The troublesome nature of the composite messages suggested that a way be found to handle them more efficiently. It was thus decided to split the composites into their component parts and treat each part as a separate message. This operation was performed on the message card files, and the modified message file was used in the sampling studies and in the Metric Search studies.

It was stated above that each question was translated into two queries. Only one set of queries was used in the category testing, carried out with the document file containing the composite messages. Both sets were used in the tests conducted with the modified file.

SECTION IV

DESCRIPTION OF PERFORMANCE MEASURES STUDIED

Two basic types of error are generated by a document retrieval system. The system can retrieve irrelevant documents, and it can fail to retrieve relevant documents. These errors, for convenience, will be called, respectively, E_1 and E_2. The performance measures chosen for studying the errors are defined as follows:

(a) The number of irrelevant messages retrieved per query = ω ,

(b) The fraction, per query, of the relevant messages retrieved = ρ (we call ρ the relevance ratio), and

(c) The fraction, per query, of retrieved messages that are relevant = τ .

Obviously, ω is a direct measure of E_1. It is useful in describing the efficiency, rather than the utility, of the system. (The most efficient machine is one which does absolutely nothing.) Both the utility and the efficiency of the system are measured in terms of ρ, which is an indirect measurement of E_2. On the other hand, τ measures neither E_1 nor E_2. It indirectly links these errors, and is another measure of efficiency.

An ideal retrieval system would yield mean values of ω, ρ, and τ as follows: $\overline{\omega} = 0$, $\overline{\rho} = 1$, $\overline{\tau} = 1$. (Note that $\overline{\omega} = 0$, $\overline{\rho} = 1$, implies $\overline{\tau} = 1$, and also that $\overline{\rho} = 1$, $\overline{\tau} = 1$, implies $\overline{\omega} = 0$.) Similarly, the most undesirable retrieval system would be described by the characteristic $\overline{\rho} = 0$ (which implies $\overline{\tau} = 0$), regardless of the value of $\overline{\omega}$. It can be seen, from this small discussion, that ρ and τ by themselves can be used to define the

the performance of document retrieval systems, and of the various query-processing techniques within those systems.

The basic errors E_1 and E_2 were broken down and studied in detail. The only criterion applied to the selection of the suberrors to be studied was that these errors be measurable. It is true that a question can be translated into both good and bad queries, and it is equally true that an apparently good query can produce poor results (low ρ), and, conversely, that an apparently poor query can yield good results. Thus, it is very difficult to quantitatively assess the effects of query quality on system performance, and no attempt was made in this evaluation to determine what percentages of the errors E_1 and E_2 were due to poor query formulation. The errors E_1 and E_2 were broken down as follows:

irrelevant messages are retrieved because of the

(a) bad synonym or subordinate relations (Class A),

(b) proper combination of, query words in single topic message, but message not relevant (Class B), or

(c) combination of query words from different parts of multitopic message (Class C);

relevant messages are missed because of the

(d) absence of one or more query words in message (Class D),

(e) misspelled query word(s) in message (Class E)*, or

(f) incomplete synonym group(s) (Class F).

*Most of the misspelled key words were put in the thesaurus as synonyms of the corresponding properly spelled key words.

The primary reasons for retrieving irrelevant messages are cited in (b) and (c). Similarly, (d) is the primary reason for missing relevant messages. The remaining sources of error are, in a sense, inseparable from their respective primary sources, but they can still be studied independently. For example, suppose that we have processed a query and have retrieved five irrelevant messages. Suppose, further, that one of the query words contains an inappropriate word in its synonym group, e.g., New York as a synonym for U.S.A. If one of the five irrelevant messages retrieved contains the word New York and not U.S.A., then the retrieval of that message will be counted as a Class A error. If three of the five irrelevant messages contain U.S.A. and not New York, and there are no other bad synonym or subordinate relations with respect to the query words, then the retrieval of those three messages will be counted as a Class B error. If the remaining message contains both New York and U.S.A., then, obviously, we cannot count the retrieval of that message exclusively as either a Class A or a Class B error. In this case, we shall regard the retrieval as a Class A and B (abbreviated $A \wedge B$) error. Many of the errors can be combined in this way. The errors are measured by simply counting, for each query with respect to a fixed query-processing method, the number of irrelevant messages retrieved, or relevant messages missed, that fall into each class. The effect of misspelled key words, of the presence of composite messages in the message file, and the quality of the thesaurus, can thus be directly measured. The two inherent errors in the operation of any coordinate indexing system, represented by Classes B and D, can also be separately examined.

Below are lists of the error combinations studied.

Irrelevant messages retrieved: A, B, C, $A \wedge B$, $A \wedge C$.

Relevant messages missed: D, E, F, $D \wedge E$, $D \wedge F$, $E \wedge F$, $D \wedge E \wedge F$.

A little reflection will show that the two combinations, B∧C and A∧B∧C, make no sense.

The means of the query-processing times for the No Expansion, Synonym Expansion, and Complete Expansion operations were acquired. The studies involving the Delete, Bibliography, and Metric Search operations were not performed on-line, and only rough estimates will be given for the additional processing time involved in each case.

Two auxiliary performance measures were derived from the basic ones. The probabilities of ω and ρ exceeding, or not exceeding, certain values were computed from histograms of the frequency of occurrence of these measures.

SECTION V

METHODS OF ANALYSIS

The basic tools used to study ROUT were statistical. Analysis of variance was applied to some phases of the experiment, e.g., category testing. Means of performance measures were arranged in rectangular tables, columns representing one source of variability in the data, and rows representing the other source.* The usual hypothesis that there is no significant difference between row (or column) means was examined at both the 1- and 5-percent significance levels.

The Mann-Whitney test is a very sensitive test used to determine the differences between two populations. The test was used to re-examine some of the inconclusive evidence derived from analyses of variance applied to both the sampling and category phases of the testing. It was used extensively to study the probabilities that ρ and ω exceed, or fail to exceed, certain values. The categories were combined in many ways in order to test for anomalies due to individual categories. Conclusions were again drawn at the 1- and 5-percent significance levels.

The chi-square test was used only once—to test for differences between the pairs of people who made up the queries and evaluated the messages with respect to the corresponding questions.

---

*There was no need to employ an analysis of variance of dimension higher than two.

Frequency of occurrence histograms were created for some of the processes with respect to the three basic performance measures ω, ρ, and τ. Some success was achieved in obtaining generalized curves for the measures ω and ρ.

Fisher's Z-test, for determining significant difference between variances, was applied mostly to the data acquired by random sampling.

A few extrapolations of system performance are made, assuming that certain error sources are either eliminated or their effects diminished.

The remainder of this section will be devoted to a derivation of the formula for computing the probability that a random selection of messages will yield a relevance ratio equal to, or greater than, a given value.

We define the following:

X = the number of messages in a message file,

I = the number of messages irrelevant to some question, and

R = the number of messages relevant to the same question.

We suppose that the question has been translated into a query and the query is processed by some kind of document retrieval system. We further define:

Y = the number of messages retrieved,

S = the number of irrelevant messages retrieved, and

T = the number of relevant messages retrieved.

Our question is now: What is the probability that a relevance ratio of at least T/R can be achieved by picking, at random, Y(= S + T) messages from the file? There are $\binom{X}{Y}$ combinations of X messages, taken Y at a time. There are $\binom{R}{T}$ ways of choosing exactly T relevant messages and $\binom{I}{S}$ ways

of choosing exactly S irrelevant messages. Hence, there are $\binom{R}{T}\binom{I}{S}$ ways of obtaining T relevant messages and S irrelevant messages by choosing Y messages from the file. Thus, the probability that a relevance ratio of exactly T/R will be achieved by picking Y messages at random is

$$\underline{P}\left(\rho = \frac{T}{R}\right) = \frac{\binom{R}{T}\binom{I}{S}}{\binom{X}{Y}} \quad ,$$

and the probability, under the same conditions, of achieving a relevance ratio of at least T/R is

$$\underline{P}\left(\rho \geq \frac{T}{R}\right) = \frac{1}{\binom{X}{Y}} \sum_{i=0}^{\min(S,\ R-T)} \binom{R}{T+i}\binom{I}{S-i} \quad . \qquad (1)$$

The upper limit of i is derived from two trivial considerations:

(a) there cannot be more relevant messages retrieved than there are messages retrieved, and

(b) there cannot be more relevant messages retrieved than there are relevant messages in the file.

Briefly, $T \leq Y$, and $T \leq R$.

After expansion and simplification, Formula (1) becomes

$$\underline{P}\left(\rho \geq \frac{T}{R}\right) =$$

$$\frac{R!\ I!\ Y!\ (X-Y)!}{X!} \sum_{i=0}^{\min(S,\ R-T)} \left[(T+i)!\ (R-(T+i))!\ (S-i)!\ (I-(S-i))!\right]^{-1} \quad . \qquad (2)$$

Probability was plotted against S for various values of T, with reference to fixed triples (X, R, I). The performances of several query processing methods are compared to the performance of random selection as exhibited in these plots.

SECTION VI

RESULTS

The ROUT evaluation was an evolutionary study in the sense that the results of one phase of the evaluation influenced the design of subsequent phases, sometimes leading to the elimination of experiments originally planned, and sometimes suggesting a deeper examination of results already obtained. The fact that so very little has been accomplished in the evaluation of document retrieval systems dictated a completely flexible mode of operation in evaluating ROUT.

The results will be presented in roughly the chronological order in which they were acquired. The experiments will be briefly described and the results will be discussed only in terms of the way they influence subsequent phases of the study. A full discussion of the results is contained in the next section.

The first phase of the testing made use of the message file containing multitopic messages (in the ratio of one multitopic message to eight single-topic messages). The messages were divided into eight mutually exclusive categories, which we shall label Cat. 0, Cat. 1, etc. Three query-processing techniques, No Expansion, Synonym Expansion, and Complete Expansion, were compared in this phase. A total of 94 questions were made up for these categories, and, for each of 82 of the questions, there was at least one relevant message in the file. Table 1 gives the frequency of relevant messages/question for each of the categories. The queries range in complexity from single-term queries to queries with five (Boolean) multiplicative factors. Table 2 gives the frequency of key word factors/query for each category.

Table 1

Relevant Messages/Question — Frequencies by Category

No. of Relevant Messages	Cat. 0	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Totals
0.5 - 2.5	5	3	4	4	3	2	5	2	28
2.5 - 4.5	1	4	2	1	1	2	2	2	15
4.5 - 6.5	1	1	1	1	2		1	2	9
6.5 - 8.5	1	3			1			1	6
8.5 - 10.5	1		1			1	1		4
10.5 - 12.5	1	1				1		2	5
12.5 - 14.5	1						1	2	4
14.5 - 16.5	1					2			3
16.5 - 18.5		1				1		1	3
18.5 - 20.5					1	1			2
>20.5						3			3
	12	13	8	6	8	13	10	11	82

Table 2

Multiplicative Factors/Query — Frequency by Category

No. of Factors	Cat. 0	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	Totals
1	2	3	1	2	1	2	1	8	20
2	3	3	1	1	2	1	3	1	15
3	5	5	4	3	2	3	4	3	29
4	2	1	1		3	4	2		13
5		1	1			3			5
	12	13	8	6	8	13	10	11	82

Table 1 shows that the majority of the questions was each answered by only a few messages. Table 2 shows the tendency of the three querists to translate the questions into simple, rather than complex, Boolean combinations. Table 3 gives means of relevance ratios for the different factor groups. The terms NE and CE are, respectively, designations for the No Expansion and Complete Expansion processes.

Table 3

Means of Relevance Ratios for Different Query Factor Groups

No. of Factors	1		2		3		4		5	
Query Processes	NE	CE	NE	CE	NE	CE	NE	CE	NE	CE
Means of ρ	0.82	0.94	0.61	0.78	0.26	0.66	0.21	0.34	0.10	0.38

From Table 3, it might be supposed that the most efficient way of using a coordinate indexing system for document retrieval is to query with only one key word, logically adding in all of its synonyms and subordinates. The strength

of this assumption does, however, depend critically upon the distribution of documents among key words, as this distribution will effect the quantity of irrelevant material retrieved. A detailed study of this phenomenon was not attempted in the ROUT evaluation, and we make these remarks only to emphasize the necessity of carefully defining the conditions under which statements about query formulation, and its effect upon retrieval, are made. Nevertheless, it is worthwhile noting the general increase in ρ as the number of multiplicative factors in the query decreases.

Tables 4 and 5 give detailed accounts of the reasons for retrieving irrelevant messages, and for failing to retrieve relevant messages. The table entries are means of errors compiled for all the categories, collectively. Table 4 contains data for all 94 of the queries. Table 5 contains data for the 74 multiple-factor queries. The terms NE, SE, and CE stand for No Expansion, Synonym Expansion, and Complete Expansion, respectively. All of the data was collected under restricted conditions. These conditions are defined by inequalities as follows:

$R > 0$ — data collected for queries whose corresponding questions are answered by relevant messages in the message file,

$\omega > 0$ — data collected for queries which yield irrelevant messages,

$\omega \geq 0$ — data collected for queries, regardless of whether they yield irrelevant messages, and

$\alpha > 0$ — data collected for queries for which relevant messages were missed.

Blocks are shaded to indicate that no entries can logically be made.

The data in the blocks defined by dotted lines are the projection of ω errors, assuming that Class C errors do not exist, as would occur in a retrieval system operating on a message file with no multitopic messages.

Table 6 contains a breakdown, by category of the ω error over all of the queries. Fisher's Z-test, for testing the significance of difference between two variances, was applied to each category for the Synonym Expansion and Complete Expansion processes. The null hypothesis is that there is no significant difference between the variances examined. The results are shown in Table 6(a).

Table 4

Error Analysis by Categories — All Queries (Entries Are Means)

Error Class	Query Process			Restrictions
	NE	SE	CE	
A		2.09	2.26	$\omega > 0$
B	5.74	3.03	4.92	$\omega > 0$
C	1.35	2.46	5.36	$\omega > 0$
A∧B		0.30	0.57	$\omega > 0$
A∧C		0.13	0.22	$\omega > 0$
D	6.15	5.02	2.26	$R > 0, \alpha > 0$
E	0.10	0.10	0.10	$R > 0, \alpha > 0$
F		0.06	0.07	$R > 0, \alpha > 0$
D∧E	0	0	0	$R > 0, \alpha > 0$
D∧F		0.40	2.33	$R > 0, \alpha > 0$
E∧F		0.02	0.02	$R > 0, \alpha > 0$
D∧E∧F	0	0	0	$R > 0, \alpha > 0$
ω	7.09	8.01	13.33	$\omega > 0$
ω	2.96(σ_ω = 7.99)	4.73(σ_ω = 9.16)	10.30(σ_ω = 14.6)	$\omega \geq 0$
ρ	0.44(σ_ρ = 0.42)	0.56(σ_ρ = 0.42)	0.71(σ_ρ = 0.37)	$R > 0$
τ	0.60(σ_τ = 0.40)	0.54(σ_τ = 0.37)	0.46(σ_τ = 0.35)	$R > 0, \omega + R - \alpha > 0$
Retrieval Time (sec)	38.2	59.1	193.8	

Table 5

Error Analysis by Categories — Multiple-Factor Queries (Entries Are Means)

Error Class	Query Process			Restrictions
	NE	SE	CE	
A		0.78	1.32	$\omega > 0$
B	3.00	2.53	4.95	$\omega > 0$
C	1.83	3.07	6.58	$\omega > 0$
A^B		0.27	0.59	$\omega > 0$
A^C		Negligible	Negligible	$\omega > 0$
D	6.40	6.02	2.42	$R > 0$, $\alpha > 0$
E	0.08	0.07	0.05	$R > 0$, $\alpha > 0$
F		Negligible	Negligible	$R > 0$, $\alpha > 0$
D^E	0	0	0	$R > 0$, $\alpha > 0$
D^F		0.51	2.57	$R > 0$, $\alpha > 0$
E^F		Negligible	Negligible	$R > 0$, $\alpha > 0$
D^E^F	0	0	0	$R > 0$, $\alpha > 0$
ω	4.83	6.75	13.07	$\omega > 0$
ω	1.87	4.05	10.45	$\omega \geq 0$
ρ	0.33	0.47	0.64	$R > 0$
Retrieval Time (sec)	39.7	65.9	232.0	
ω - No Class C errors	3.00	3.68	6.49	$\omega > 0$

Table 6

Means and Standard Deviations of ω by Categories — All Queries

Category	Size	Query Process			Mean / Standard Deviation
		NE	SE	CE	
0	14	4.21 10.56	7.5 12.38	10.36 12.28	$\bar{\omega}$ / σ_ω
1	14	0.86 1.35	1.29 2.12	6.07 9.58	$\bar{\omega}$ / σ_ω
2	10	1.7 2.76	2.2 2.89	7.6 5.99	$\bar{\omega}$ / σ_ω
3	8	6.75 17.86	7.5 17.65	15.38 18.30	$\bar{\omega}$ / σ_ω
4	10	1.1 1.22	5.3 8.28	14.3 23.22	$\bar{\omega}$ / σ_ω
5	13	2.15 3.78	3.77 5.73	11.46 15.29	$\bar{\omega}$ / σ_ω
6	12	3.58 5.44	6.25 7.39	11.75 10.44	$\bar{\omega}$ / σ_ω
7	13	4.08 8.53	4.85 8.29	8.15 14.66	$\bar{\omega}$ / σ_ω

Table 6(a)

Analysis of Variance Results

Category	Hypothesis Rejected at	
	5% level	1% level
0	No	No
1	Yes	Yes
2	Yes	No
3	No	No
4	Yes	Yes
5	Yes	Yes
6	No	No
7	Yes	No

The conclusion here is that there are significant variations due to both categorization and the method of query processing. More specifically, the effects due to categorization are more significant at the 5-percent level than at the 1-percent level. An analysis of variance (run for all three query-processing methods) reveals the hypotheses (a) that there is no significant difference between row means and (b) no significant difference between column means, must both be rejected at both the 5- and 1-percent levels.

Tables 7 and 8 are, respectively, breakdowns by category of the ρ and τ errors over all of the queries. Following each table (6, 7, and 8) are the results of an analysis of variance carried out on the means entered in the table.

Table 7

Means and Standard Deviations of ρ by Categories — All Queries (R > 0)

Category	Size	Query Process			Mean / Standard Deviation
		NE	SE	CE	
0	12	0.48 0.41	0.68 0.31	0.75 0.23	$\bar{\rho}$ / σ_ρ
1	13	0.46 0.38	0.53 0.42	0.79 0.30	$\bar{\rho}$ / σ_ρ
2	8	0.33 0.44	0.46 0.47	0.73 0.33	$\bar{\rho}$ / σ_ρ
3	6	0.28 0.37	0.36 0.29	0.36 0.29	$\bar{\rho}$ / σ_ρ
4	8	0.43 0.46	0.56 0.45	0.81 0.26	$\bar{\rho}$ / σ_ρ
5	13	0.28 0.34	0.39 0.39	0.48 0.42	$\bar{\rho}$ / σ_ρ
6	10	0.43 0.41	0.50 0.41	0.67 0.39	$\bar{\rho}$ / σ_ρ
7	12	0.75 0.34	0.88 0.15	0.94 0.11	$\bar{\rho}$ / σ_ρ

Fisher's Z-test was again applied to each category for the Synonym Expansion and Complete Expansion processes. The null hypothesis is that there is no significant difference between the variances examined. The results are that the hypothesis is not rejected at either the 5- or 1-percent level for any

category. Thus, we proceed to an analysis of variance of the means for all three query processes (Table 7(a)). The analysis reveals the hypotheses (a) that there is no significant difference between row means, and (b) no significant difference between column means, must both be rejected at both the 5- and 1-percent levels. Five subanalyses were run to determine which categories were making it impossible to accept the hypothesis that there is no significant difference in the row means. The results are:

(a)	All categories except No. 3:	both hypotheses rejected at both the 5- and 1-percent levels.
(b)	All categories except No. 3 and No. 5:	both hypotheses rejected at both the 5- and 1-percent levels.
(c)	All categories except No. 3, 5 and 7:	reject hypothesis of equal column means at the 5- and 1-percent levels. Do not reject, at either the 5- or 1-percent level, hypothesis of equal row means.
(d)	All categories except No. 7:	both hypotheses rejected at both the 5- and 1-percent levels.
(e)	All categories except No. 3 and 7:	both hypotheses reject at both the 5- and 1-percent levels.

One final analysis of variance was run for all categories, but only the relevance ratio means was used for the No Expansion and Complete Expansion processes. It was again impossible, at either the 5- or 1-percent level of significance, to accept the hypothesis of equal row means. A complete discussion of these results, and what they imply, will be found in the next section.

Table 7(a)

Means of ρ by Categories — Multiple-Factor Queries

Category	Query Process		
	NE	SE	CE
0	0.37	0.61	0.71
1	0.41	0.51	0.75
2	0.28	0.42	0.74
3	0.04	0.17	0.17
4	0.34	0.50	0.78
5	0.17	0.28	0.38
6	0.36	0.44	0.63
7	0.66	0.68	0.89

Table 8

Means and Standard Deviations of τ by Categories — All Queries
($R > 0$, $\omega + R - \alpha > 0$)

Category	Size	Query Progress			Mean / Standard Deviation
		NE	SE	CE	
0	7, 11, 11	0.59 0.32	0.56 0.31	0.44 0.26	$\overline{\tau}$ / σ_τ
1	9, 10, 12	0.81 0.21	0.74 0.21	0.54 0.30	$\overline{\tau}$ / σ_τ
2	5, 5, 7	0.31 0.37	0.35 0.34	0.34 0.29	$\overline{\tau}$ / σ_τ
3	3, 4, 5	0.68 0.45	0.51 0.49	0.41 0.48	$\overline{\tau}$ / σ_τ
4	8, 8, 8	0.50 0.50	0.52 0.44	0.56 0.40	$\overline{\tau}$ / σ_τ
5	9, 10, 13	0.62 0.41	0.57 0.36	0.43 0.38	$\overline{\tau}$ / σ_τ
6	9, 10, 10	0.39 0.35	0.20 0.16	0.16 0.11	$\overline{\tau}$ / σ_τ
7	11, 12, 12	0.77 0.30	0.72 0.31	0.67 0.32	$\overline{\tau}$ / σ_τ

An analysis of variance carried out on the table means revealed that the hypotheses of equal column and equal row means cannot be rejected at either the 5- or 1-percent level of significance. The conclusion is that $\overline{\tau}$ is not influenced by any of the three query-processing methods involved. This is a reasonable conclusion, since increases in the relevance ratio are accompanied by increases in ω for the three processes.

Figures 1 through 12 present frequency of occurrence histograms for ρ and ω.

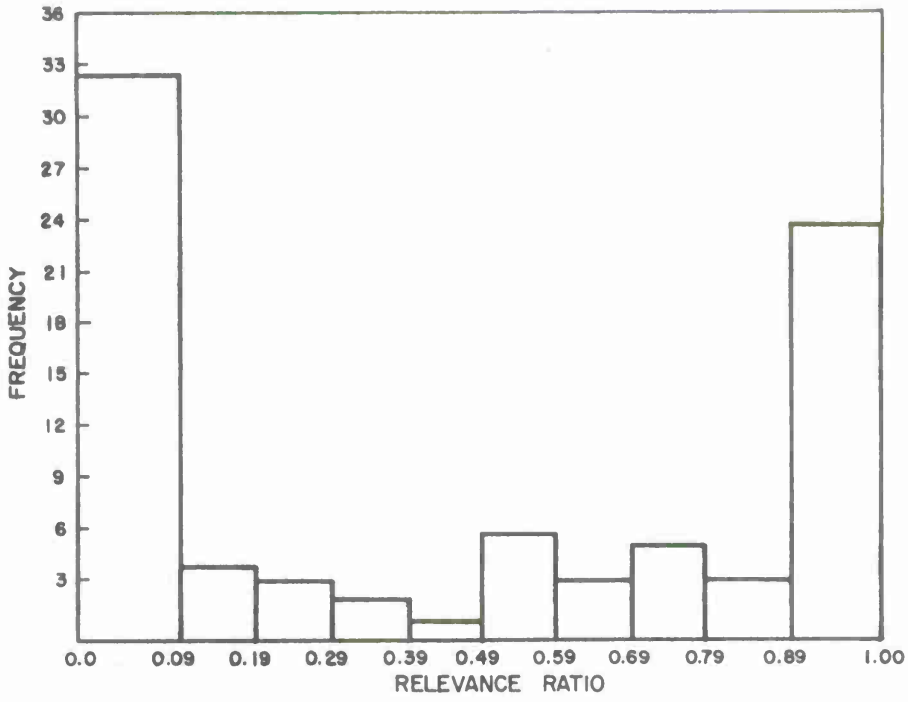


Fig. 1. Frequency of Occurrence of Relevant Messages Retrieved — All Queries, No Expansion ($R \neq 0$)

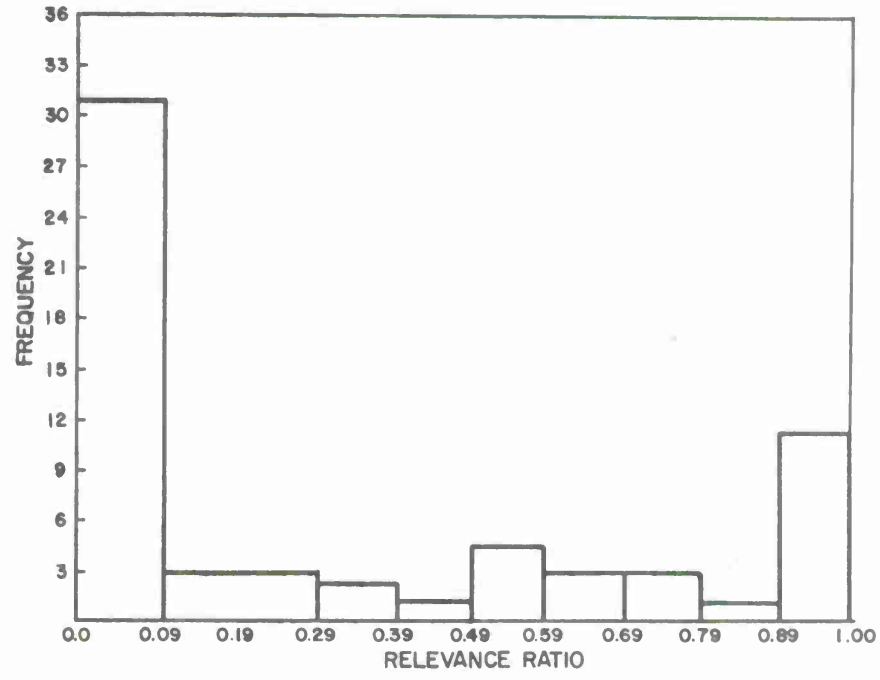


Fig. 2. Frequency of Occurrence of Relevant Messages Retrieved — Multiple Factor Queries, No Expansion ($R \neq 0$)

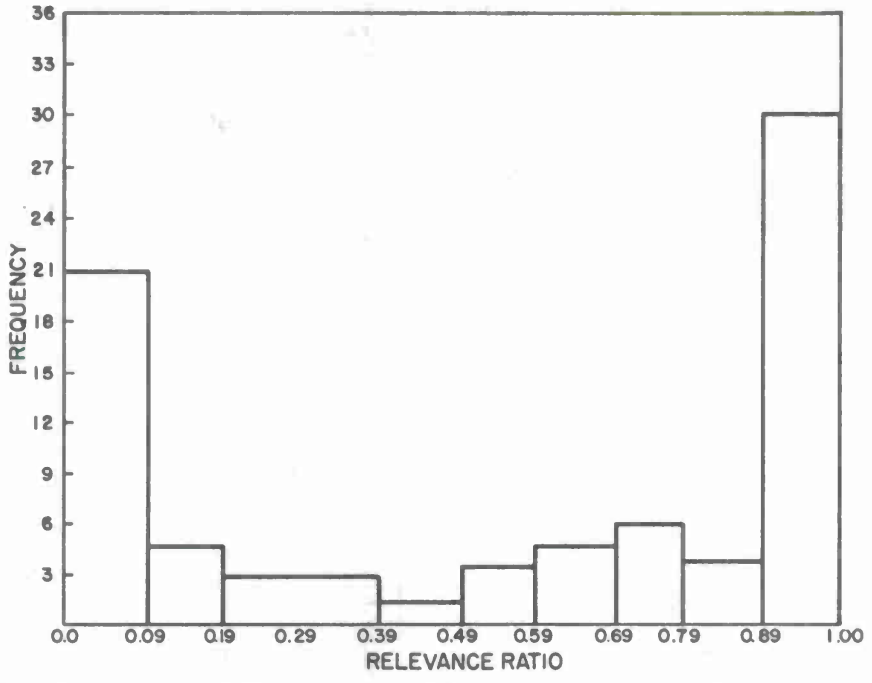


Fig. 3. Frequency of Occurrence of Relevant Messages Retrieved — All Queries, Synonym Expansion (R≠0)

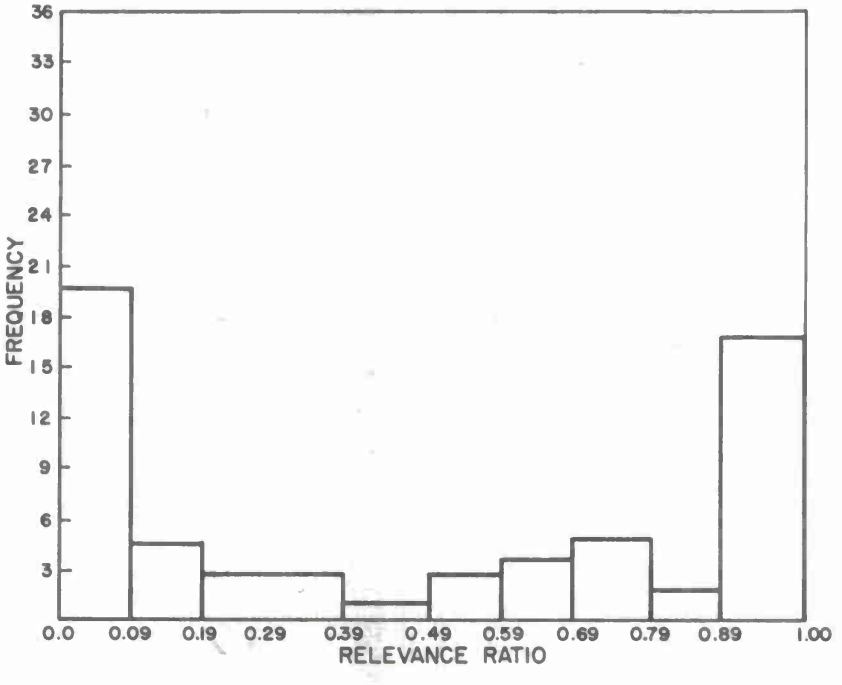


Fig. 4. Frequency of Occurrence of Relevant Messages Retrieved — Multiple Factor Queries, Synonym Expansion (R≠0)

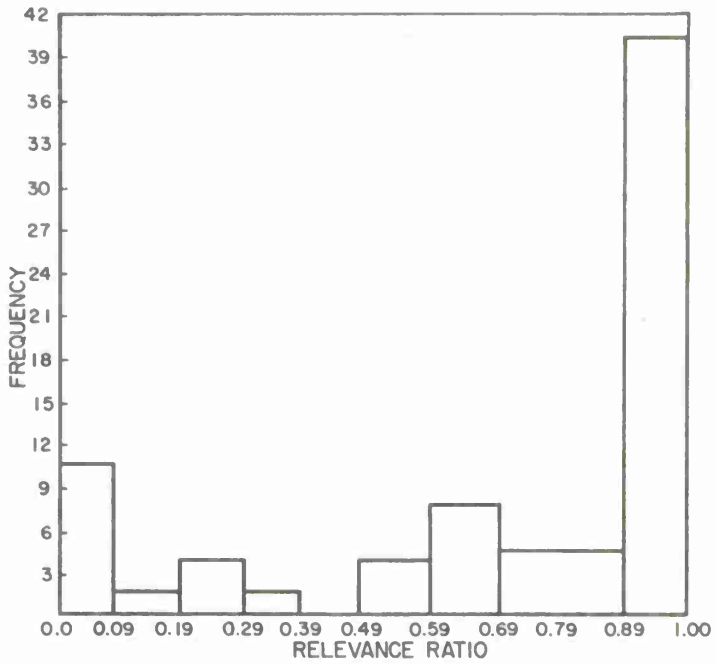


Fig. 5. Frequency of Occurrence of Relevant Messages Retrieved — All Queries, Complete Expansion (R≠0)

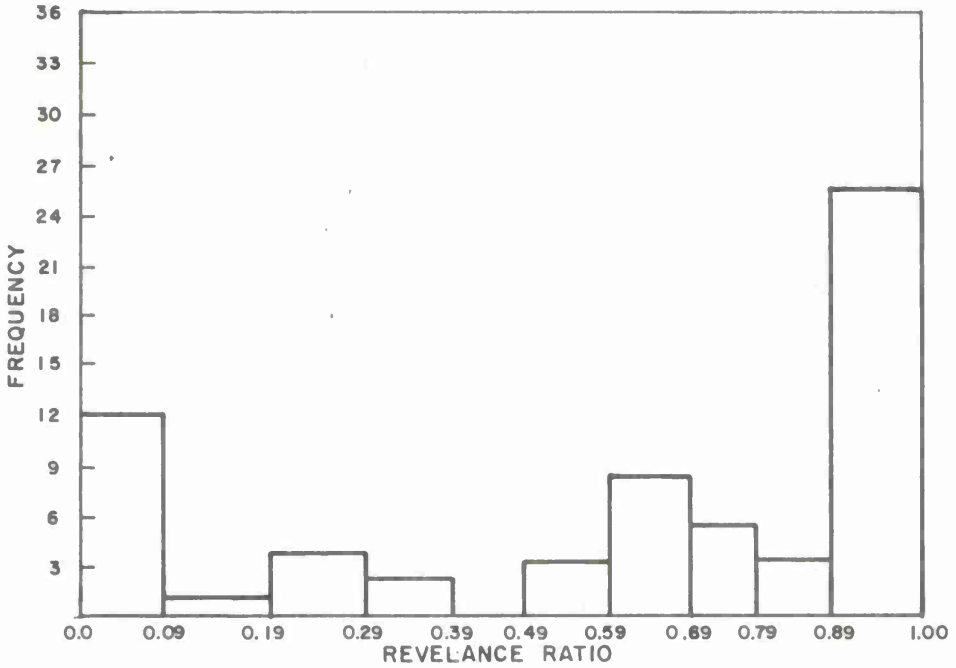


Fig. 6. Frequency of Occurrence of Relevant Messages Retrieved — Multiple Factor Queries, Complete Expansion (R≠0)

Parabolas were fitted to the histograms in Figs. 2, 4 and 6. Their equations are as follows:

(a) $F = 72(\rho^2 - 1.28_{\rho} + 0.43)$ No Expansion,

(b) $F = 56(\rho^2 - 1.05_{\rho} + 0.36)$ Synonym Expansion, and

(c) $F = 63(\rho^2 - 0.8_{\rho} + 0.19)$ Complete Expansion.

These curves are interesting because of the regularity in the changes of the constants as expansion capability is added.

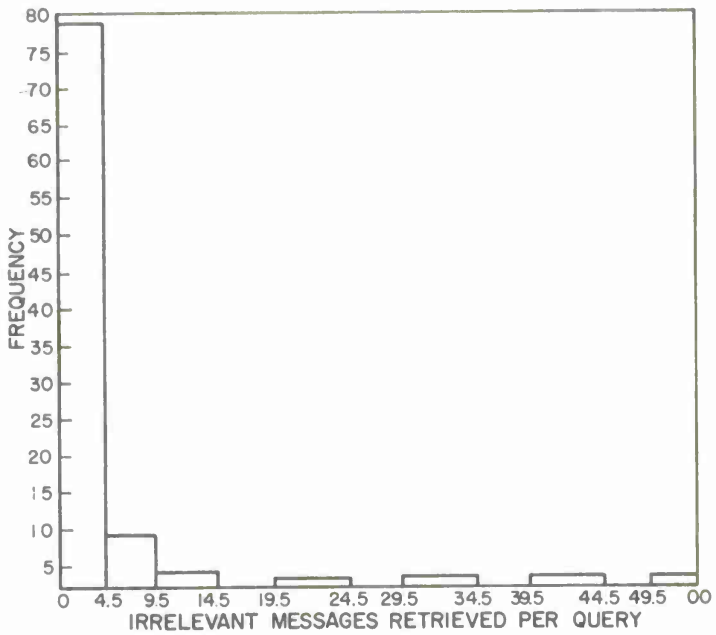


Fig. 7. Frequency of Occurrence of Irrelevant Messages Retrieved — All Queries, No Expansion

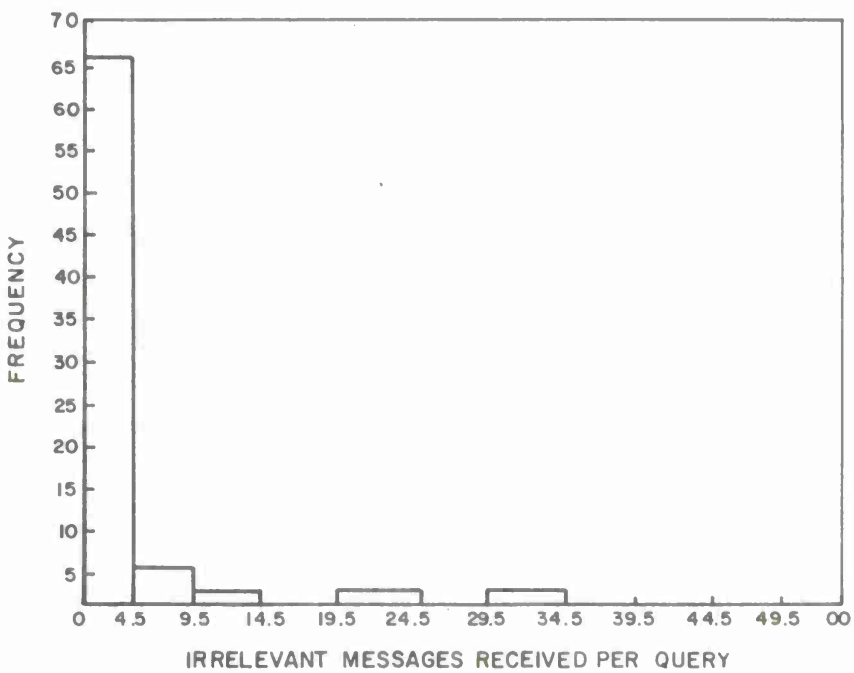


Fig. 8. Frequency of Occurrence of Irrelevant Messages Retrieved — Multiple Factor Queries, No Expansion

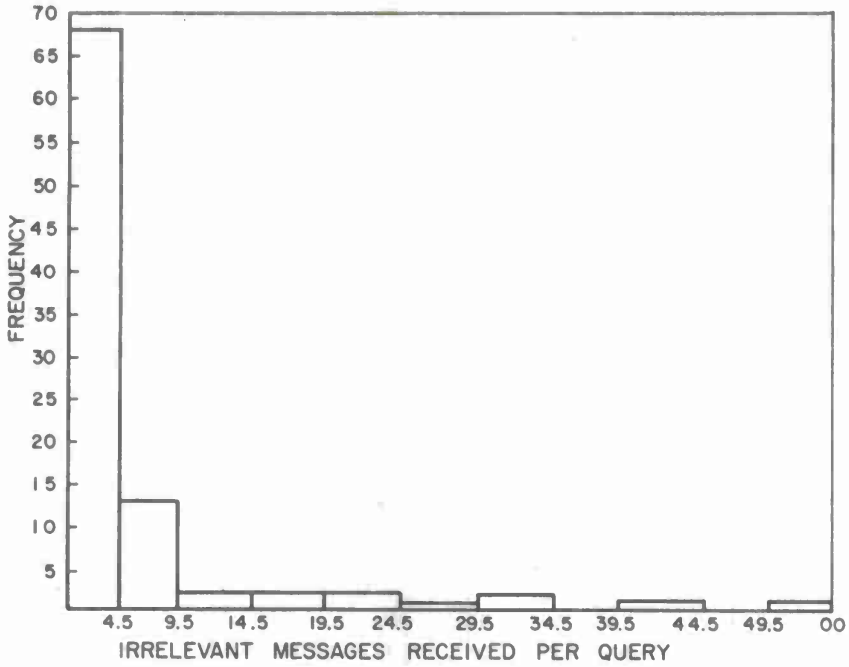


Fig. 9. Frequency of Occurrence of Irrelevant Messages Retrieved — All Queries, Synonym Expansion

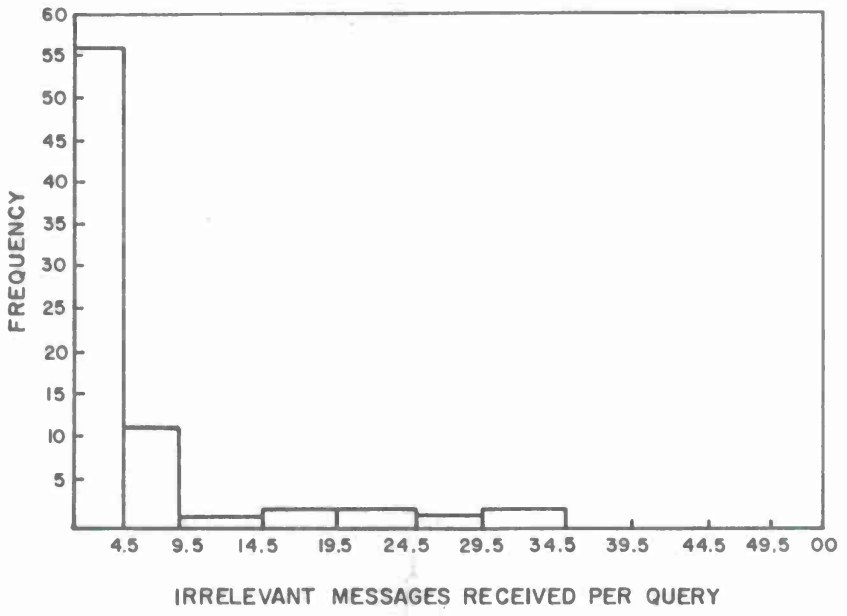


Fig. 10. Frequency of Occurrence of Irrelevant Messages Retrieved — Multiple Factor Queries, Synonym Expansion

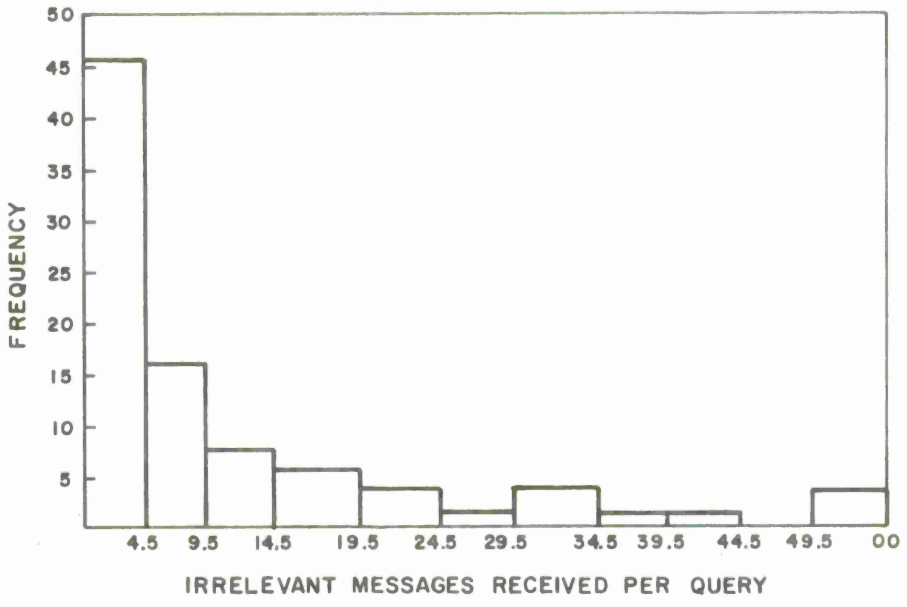


Fig. 11. Frequency of Occurrence of Irrelevant Messages Retrieved — All Queries, Complete Expansion

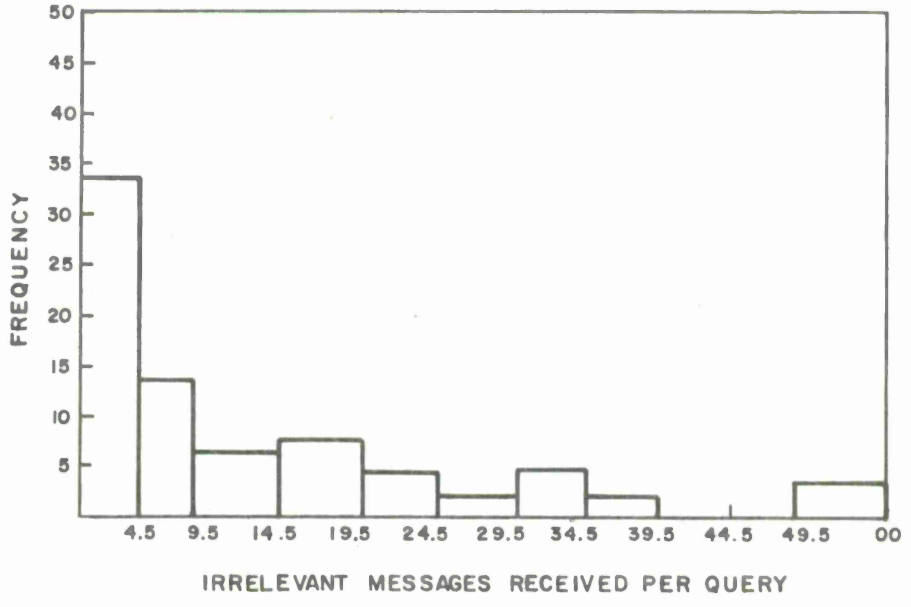


Fig. 12. Frequency of Occurrence of Irrelevant Messages Retrieved — Multiple Factor Queries, Complete Expansion

Table 9

Probability of ρ Exceeding, or not Exceeding, Given Values (R > 0)

63 Queries	NE	SE	CE
$\underset{\pm}{P}$ ($\rho < 0.5$)	0.63	0.51	0.30
$\underset{\pm}{P}$ ($\rho < 0.4$)	0.62	0.49	0.30
$\underset{\pm}{P}$ ($\rho < 0.3$)	0.59	0.45	0.27
$\underset{\pm}{P}$ ($\rho < 0.2$)	0.54	0.40	0.21
$\underset{\pm}{P}$ ($\rho < 0.1$)	0.49	0.32	0.19
$\underset{\pm}{P}$ ($\rho \geq 0.5$)	0.36	0.49	0.70
$\underset{\pm}{P}$ ($\rho \geq 0.6$)	0.29	0.45	0.65
$\underset{\pm}{P}$ ($\rho \geq 0.7$)	0.24	0.38	0.52
$\underset{\pm}{P}$ ($\rho \geq 0.8$)	0.19	0.30	0.45
$\underset{\pm}{P}$ ($\rho \geq 0.9$)	0.17	0.27	0.40

Table 10

Probability of ω Exceeding, or not Exceeding, Given Values

75 Queries	NE	SE	CE
$\underset{\pm}{P}\ (\omega \leq 4)$	0.88	0.75	0.45
$\underset{\pm}{P}\ (\omega \leq 9)$	0.96	0.89	0.63
$\underset{\pm}{P}\ (\omega \leq 14)$	0.97	0.91	0.72
$\underset{\pm}{P}\ (\omega \leq 19)$	0.97	0.93	0.83
$\underset{\pm}{P}\ (\omega \leq 24)$	0.99	0.96	0.91
$\underset{\pm}{P}\ (\omega > 4)$	0.12	0.25	0.55
$\underset{\pm}{P}\ (\omega > 9)$	0.04	0.11	0.37
$\underset{\pm}{P}\ (\omega > 14)$	0.03	0.09	0.28

The Mann-Whitney test was applied to the probability data in two different ways. In one case, eight combinations of four categories each were formed, and the probabilities above were computed for each of the combinations. The Mann-Whitney test was then applied for each of the parenthetic values given in Tables 10 and 11. The result was rejection of the hypothesis of equal $\underset{\pm}{P}$-values among each of the pairs of query processes at both the 5- and 1-percent levels of significance. The same kind of test was applied to eight combinations of two categories each. The result, in this case, was almost universal rejection of the hypothesis of equal $\underset{\pm}{P}$-values with respect to the relevance ratio at the 5-percent level of significance, and, similarly, almost universal nonrejection of the hypothesis at the 1-percent level for the No Expansion — Synonym Expansion pair of query processes. The conclusions are that categorization does have an

effect on the $P_{\pm}$-values, and that the effect of categorization disappears when the categories are collectively studied in groups of four. Significant differences at the 5- and 1-percent levels were again observed between the query-processing methods.

The time (in seconds) required for retrieval is given in Table 11. The limits of the time interval are defined by the inputting of the query and the completion of the printout of the retrieved messages. (Queries were punched on cards and batch-processed.)

Table 11

Times for Retrieval (Means)

Category	Query Process		
	NE	SE	CE
0	32.6	57.0	186.4
1	40.2	54.9	146.8
2	33.8	50.9	122.5
3	30.3	38.0	186.4
4	32.5	66.8	348.0
5	46.1	73.7	311.4
6	44.5	74.9	209.0
7	39.2	46.8	101.1

An analysis of variance on the means of Table 11 revealed no significant difference between the categories at either the 5- or 1-percent level, whereas the hypothesis of equal column means was rejected at both the 5- and 1-percent levels of significance.

Table 12 is the basis for an X^2 - test to determine if there are significant differences between the pairs of people rating the relevancy of messages to the questions they asked of the system, and those formulating the queries for the questions. The performance measure used is the mean of the relevance ratio, and the eight pairs correspond one-to-one to the eight categories. The relevance ratios are over all queries for the Complete Expansion query process.

Table 12

Message Raters Versus Query Formulators

		Person Determining Relevancy	
Person Formulating Queries		D. Budde	J. Rial
		0.361	
	D. Budde	0.476	
		0.940	
	R. Radner	0.790	0.728
		0.811	0.666
	J. Rial		0.752

The test resulted in no rejection of the hypothesis of equal relevance ratio means among the pairs.

The next phase of the testing involved sampling the message file in both random and semirandom modes. The predictable and rather uninteresting effects of using single-factor queries were investigated in the category phase of the evaluation, and it was decided to exclude these queries from further testing. The remaining (multiple-factor) queries of the category testing were re-used along with a new batch of queries, one for each of the questions

originally translated into a multiple-factor query. The original queries are designated as Set A, and the new batch as Set B. The undesirable effects of having multitopic messages in the file was also studied in the category testing. The new file was, consequently, almost entirely purged of such messages by breaking them up and treating each part as an individual message.

The expansion of a query can never have the effect of decreasing either ρ or ω. The Delete and Bibliography processes can, on the other hand, never increase either ρ or ω. The performance of the two expansion processes was studied in the category testing, and, once again, in the sampling studies. This made possible a comparison of the techniques of gathering the data. The Bibliography and Delete processes were studied singly and in combination in the sampling tests, and also in the more extensive tests which followed.

Very early in the data-gathering stage of the sampling tests, it became clear that the detailed error analysis carried out in the category study could not be made by sampling. The thinness of the data simply did not permit the forming of meaningful conclusions. The three performance measures, ρ, ω, and τ, were consequently the only ones examined in the sampling study.

Table 13 gives the means and standard deviations of the relevance ratio for eight query processes. The truly random sample is denoted by TR. The other two samples, each consisting of 100 consecutive messages, are denoted C_1 and C_2. The Bibliography and Delete processes are denoted, respectively, by BIB and DEL. The terms A and B refer to the two query sets described above.

Table 13

Relevance Ratios — Samples of 100 Messages, 77-Multiple Factor Queries (R > 0)

Sample	Query Set	Query Process								Mean / Std. Dev.
		NE, BIB	SE, BIB	CE, BIB	CE, BIB, DEL	NE	SE	CE	CE, DEL	
TR	A	0.17 0.38	0.29 0.45	0.50 0.47	0.35 0.45	0.21 0.41	0.33 0.46	0.55 0.48	0.41 0.47	$\bar{\rho}$ / σ_ρ
TR	B	0.24 0.42	0.33 0.45	0.52 0.46	0.40 0.46	0.24 0.42	0.33 0.45	0.54 0.47	0.42 0.47	$\bar{\rho}$ / σ_ρ
C_1	A	0.16 0.36	0.21 0.41	0.43 0.48	0.37 0.45	0.16 0.36	0.21 0.41	0.48 0.49	0.43 0.46	$\bar{\rho}$ / σ_ρ
C_1	B	0.30 0.41	0.40 0.44	0.54 0.47	0.52 0.45	0.30 0.41	0.40 0.44	0.62 0.47	0.58 0.45	$\bar{\rho}$ / σ_ρ
C_2	A	0.24 0.38	0.34 0.41	0.61 0.44	0.47 0.43	0.24 0.38	0.37 0.43	0.70 0.40	0.52 0.42	$\bar{\rho}$ / σ_ρ
C_2	B	0.36 0.44	0.50 0.44	0.63 0.44	0.47 0.43	0.37 0.43	0.46 0.43	0.68 0.43	0.49 0.44	$\bar{\rho}$ / σ_ρ

Fisher's Z-test was applied to the variance pairs for the Synonym Expansion and Complete Expansion processes. The result was nonrejection of the hypothesis of equal variances at both the 5- and 1-percent levels of significance. An analysis of variance was made of the relevance ratio means for the three processes, NE, SE, and CE. The result was rejection of each of the hypotheses of equal column and equal row means at both the 5- and 1-percent levels of significance. The same results were achieved when the analysis was rerun with the data for C_1-B excluded. No difference in the results occurred when the data for samples C_1-B and C_2-A were excluded. The conclusion is that sampling has a significant effect on the values of the relevance ratio.

Table 14 gives the results of applying the Mann-Whitney test to the relevance ratio means for selected pairs of query processes.

Table 14

Results of Applying Mann-Whitney Test to the Relevance Ratio Means for Selected Pairs of Query Processes

Query Process Pair	Hypothesis of Equal Means Rejected at	
	5% level	1% level
NE, BIB-NE	No	No
SE, BIB-SE	No	No
CE, BIB-CE	No	No
CE, BIB, DEL-CE, DEL	No	No
CE, BIB, DEL-CE, BIB	Yes	No
CE, DEL-CE	Yes	No

From Table 14, we conclude that the Delete process has a greater effect than the Bibliography process in reducing $\overline{\rho}$.

Table 15 gives $\overline{\omega}$ and σ_{ω} over all the queries for eight query processes.

Table 15

ω — Samples of 100 Messages, 77-Multiple Factor Queries ($\omega \geq 0$)

Sample	Query Set	Query Process								Mean / Std. Dev.
		NE, BIB	SE, BIB	CE, BIB	CE, BIB, DEL	NE	SE	CE	CE, DEL	
TR	A	0.013 0.11	0.091 0.33	0.208 0.52	0.085 0.28	0.078 0.27	0.273 0.64	1.26 3.24	0.437 0.82	$\bar{\omega}$ / σ_ω
TR	B	0.096 0.38	0.151 0.46	0.247 0.57	0.197 0.47	0.301 0.82	0.479 1.12	0.918 1.80	0.682 1.50	$\bar{\omega}$ / σ_ω
C_1	A	0 0	0.026 0.16	0.039 0.19	0.028 0.17	0.026 0.16	0.091 0.37	0.39 1.34	0.099 0.34	$\bar{\omega}$ / σ_ω
C_1	B	0.081 0.49	0.108 0.51	0.108 0.51	0.162 0.70	0.135 0.55	0.162 0.57	0.338 0.76	0.279 0.80	$\bar{\omega}$ / σ_ω
C_2	A	0.013 0.11	0.078 0.42	0.169 0.52	0.143 0.46	0.078 0.39	0.273 0.78	0.948 2.29	0.529 1.14	$\bar{\omega}$ / σ_ω
C_2	B	0.042 0.20	0.111 0.46	0.181 0.54	0.104 0.43	0.139 0.54	0.25 0.76	0.542 1.15	0.284 0.81	$\bar{\omega}$ / σ_ω

Fisher's Z-test for comparing variances hardly needs to be applied to the data in Table 15. A visual inspection of the table shows a variation in σ_ω so enormous it can only be concluded that random (or semirandom) sampling at the 100-message level cannot be used to accurately measure $\bar{\omega}$. The cause of this failure cannot be stated in terms of a difference between the two sets of queries. It does not appear to be a function of the kind of sampling employed. A closer study of the frequencies of the number of irrelevant messages retrieved, and of the number of queries which yielded irrelevant messages at all, reinforces the conclusion that the data were too thin to be representative. The Mann-Whitney test did, however, result in the rejection, at the 5-percent level of significance, of the hypothesis of equal ω means for the query process pairs CE-SE, NE-SE (therefore, CE-NE), CE, BIB-SE, BIB and NE, BIB-SE, BIB (consequently, CE, BIB-NE, BIB). Rejection at the 1-percent level occurred only for the pair CE-SE. Other results of the application of the Mann-Whitney test are shown in Table 16.

Table 16

Results of Applying Mann-Whitney Test

Query Process Pair	Hypothesis of Equal ω Means Rejected at	
	5% level	1% level
NE, BIB-NE	Yes	No
SE, BIB-SE	Yes	Yes? *
CE, BIB-CE	Yes	Yes
CE, BIB, DEL-CE, DEL	Yes	Yes ?
CE, BIB, DEL-CE, BIB	No	No
CE, DEL-CE	Yes	No

*Question marks signify borderline rejections.

From the third, fifth and sixth rows in the table, we conclude that the Bibliography process is more powerful than the Delete process in reducing $\bar{\omega}$.

The means and standard deviations of τ for the samples are given in Table 17. The data are given only for the CE, DEL and CE, DEL, BIB processes since it is abundantly clear, at this point, that the NE and SE processes are relatively ineffectual.

Table 17

Means and Standard Deviations of τ — Samples — ($R > 0$, $\omega + R - \alpha > 0$)

Sample	Query Set	Query Process		Mean / Standard Deviation
		CE, DEL, BIB	CE, DEL	
TR	A	0.75 / 0.37	0.57 / 0.41	$\bar{\tau}$ / σ_τ
TR	B	0.60 / 0.43	0.45 / 0.45	$\bar{\tau}$ / σ_τ
C_1	A	1 / 0	0.93 / 0.21	$\bar{\tau}$ / σ_τ
C_1	B	0.98 / 0.06	0.92 / 0.19	$\bar{\tau}$ / σ_τ
C_2	A	0.82 / 0.36	0.71 / 0.41	$\bar{\tau}$ / σ_τ
C_2	B	0.89 / 0.27	0.83 / 0.30	$\bar{\tau}$ / σ_τ

The Mann-Whitney test was applied to the means for the two processes. The hypothesis of equal column means could not be rejected at either the 5- or 1-percent level of significance.

A comparison was made of the relevance ratio means for the six samples and the eight categories with respect to the NE, SE, and CE query processes. The Mann-Whitney test was employed. The hypothesis of equal means could not be rejected at either the 5- or 1-percent level of significance in any of the three cases.

The histograms of the frequency of occurrence of ω and ρ values in given intervals show the same general features, relative to a given query process, as those derived from the category testing.

The last major phase of the evaluation involved the use of 52 queries which were selected so as to simultaneously satisfy the following criteria:

(a) at least one relevant message exists in the file for each question, and

(b) for each query, at least one message is chosen on the basis of the bibliographic printout of the computed reply.

Each of the questions corresponding to the 52 queries was rated against the entire message file. The mean and standard deviation of the relevance ratio was computed for six query processing methods. The results are given in Table 18.

Table 18

Relevance Ratios for Six Query Processes ($R \neq 0$)

Query Process	$\overline{\rho}$	σ_ρ
SE + BIB	0.55	0.33
SE + BIB + METRIC	0.66	0.34
SE + BIB + METRIC^2	0.68	0.33
CE + BIB	0.70	0.32
CE + BIB + METRIC	0.80	0.29
CE + BIB + METRIC^2	0.81	0.30

The Delete process was again investigated, using the same 52 queries that provided the data for Table 18; and the results are displayed in Table 19.

Table 19

Values of ρ, ω, and τ for the Delete Process

Query Process		Measure	Condition
CE + DEL	CE + DEL + BIB		
0.65	0.63	$\overline{\rho}$	$R > 0$
0.35	0.34	σ_ρ	
4.08	1.69	$\overline{\omega}$	$\omega \geq 0$
7.67	3.03	σ_ω	
0.60	0.72	$\overline{\tau}$	$R > 0$,
0.36	0.32	σ_τ	$\omega + R - \alpha > 0$

Encouraged by the results shown in Table 18, a final comparison was made between the CE + BIB and CE + BIB + METRIC processes. Both query sets were used, subject to the restrictions given in the paragraph preceding Table 18, along with the additional restriction that no query would be used if the relevance ratio equaled 1.0 for the CE + BIB process. The absolute improvement in performance afforded by the metric could thus be measured. Table 20 and Fig. 13 give the data for the 56 queries which satisfied the three conditions.

Table 20

CE + BIB Versus CE + BIB + METRIC, Relevance Ratios

Query Process	$\bar{\rho}$	σ_ρ
CE + BIB	0.43	0.65
CE + BIB + METRIC	0.60	0.28

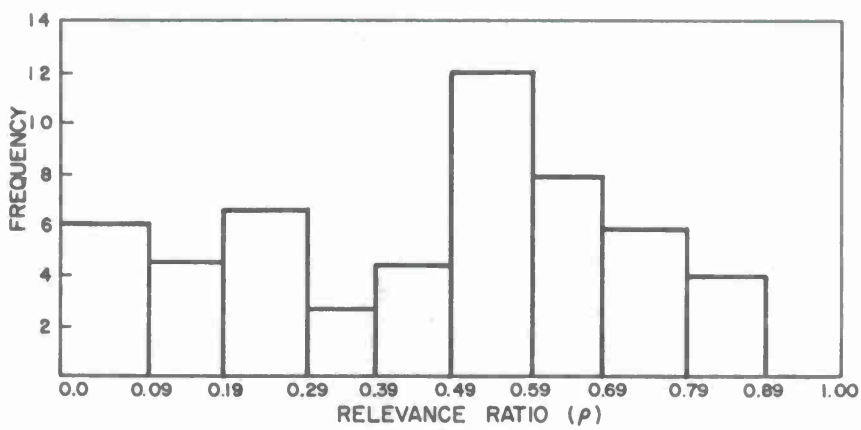


Fig. 13. Frequency of Occurrence of ρ (CE + BIB)

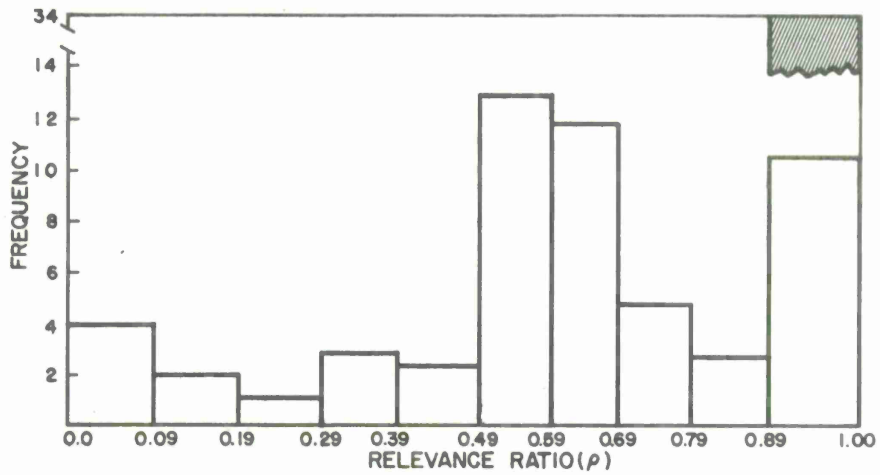


Fig. 14. Frequency of Occurrence of ρ (CE + BIB + METRIC)

The significance of Table 20 and Fig. 13 lie in the rather dramatic increase in the higher values of the relevance ratio. The shift is expressed most powerfully in the difference in the σ_ρ values for the two processes. The shaded portion of Fig. 14 represents the total number of queries which yielded a relevance ratio of unity for the CE + BIB + METRIC process, irrespective of whether the corresponding relevance ratio for the CE + BIB process was unity. Thus, the Metric Search process increased the number of perfect retrievals by eleven, and approximately one third of the total number of perfect retrievals are due to that process. The parabolic nature of the previous frequency plots is hardly to be observed in Figs. 13 and 14. The greatest divergence appears in the lower range of relevance ratio values, and this variation is possibly due to a slight difference in the rating of the messages. A more plausible explanation

is that the queries used for Figs. 13 and 14 produced bibliographies from which at least one message was judged relevant, and, thus, the likelihood of retrieving at least one relevant message per query was increased.

The distribution of message-message distances is shown in Fig. 15. The hump at $\delta \approx 0.25$ is due to a large group of messages written in a very similar manner on the same general topic. The hump at $\delta \approx 0.75$ shows that most of the messages do not have an appreciable number of key words in common with other messages. There were six pairs of near-duplicate messages in the file which contributed to the $f = 6$ point at $\delta = 0$. The absence of messages a unit distance from any other message is accounted for by the presence of addresses, originators, and classifications in every message, and the widespread distribution of these nontextual words among the messages.

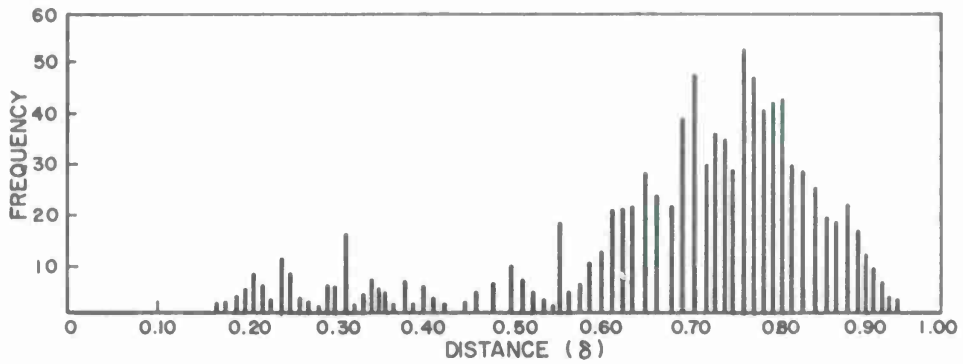


Fig. 15. Frequency of Occurrence of Message—Message Distances

The remaining material of this section is devoted to a presentation of curves describing the probability, under specific conditions, of achieving a relevance ratio equal to or greater than a given value by randomly selecting messages from the message file. Probability is plotted against the maximum number of irrelevant messages retrieved, for several values of ρ. Two hypothetical models are used, each based on a 100-message file, which is roughly two thirds the average size of a category used in the category testing. The first model supposes a division of the file into 10 relevant and 90 irrelevant messages (Fig. 16), and the second uses a division of the file into 5 relevant and 95 irrelevant messages (Fig. 17).

We shall give an example of how to read and use these graphs. From Fig. 16, we find that the probability of obtaining $\rho \geq 0.1$ by randomly selecting six messages from the file is 0.47. (At least one of the ten relevant messages is selected, and no more than five irrelevant messages are retrieved.)

A comparison of the performances of the random selection process and the No Expansion process (see Table 9) reveals that the No Expansion process, in many cases, offers only a little more capability than the crudest retrieval mechanism conceivable. Many of the divisions of the categories into relevant and irrelevant messages closely match the models upon which Fig. 16 and 17 are based, and many of the associated retrievals contain relevant and irrelevant messages in the proportions used to derive the graphs. It is considerations of this type which help make the data of Table 9 reflect the great increase in capability achieved through a thesaurus of synonym and subordinate groupings.

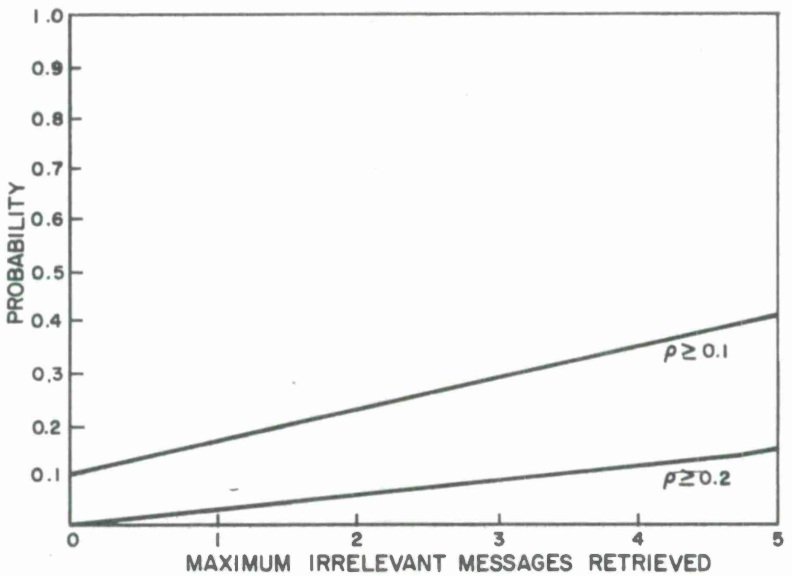


Fig. 16. Retrieval by Random Selection of Messages — 10 Relevant Messages, 90 Irrelevant Messages

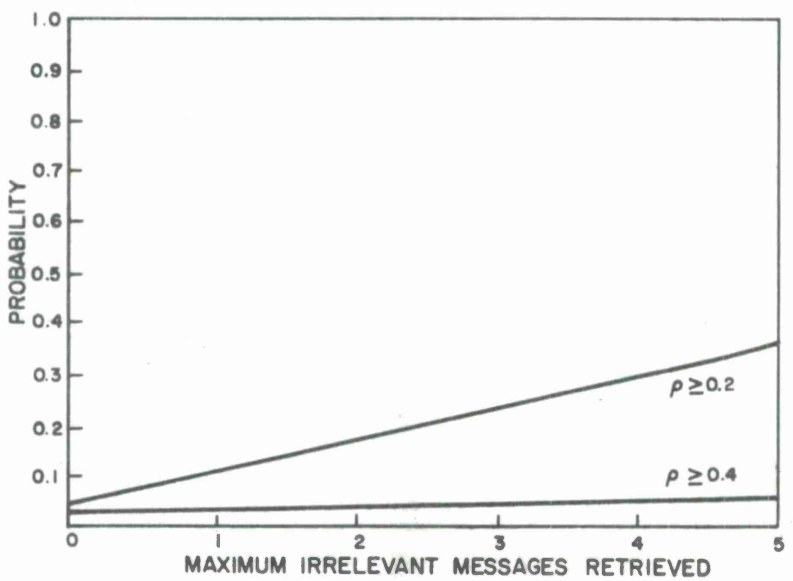


Fig. 17. Retrieval by Random Selection of Messages — 5 Relevant Messages, 95 Irrelevant Messages

SECTION VII

DISCUSSION OF RESULTS

The basic defects of coordinate indexing are most clearly represented by the data in Tables 4 and 5. Error A is seen to be less significant than the more inherent error B, and, similarly, error D, along with $D \wedge F$, is overwhelmingly more important than error F by itself. Improvements in the thesaurus will, in fact, have less influence on increasing the relevance ratio than on reducing the number of irrelevant messages retrieved. The undesirable effect of having multitopic messages in the file has been noted before, but it is worth reemphasizing. The problem of mispelled key words can easily be remedied by putting them in the synonym groups of the corresponding properly spelled words.

The effect of using single-term queries can be seen most readily by comparing Error A and the values of ω in the two tables. The multiplication of query words has the effect of reducing the number of irrelevant messages generated because of any specific expansion term, and Error A shows the magnitude of this reduction. The term ω, for the two processes, Synonym Expansion and Complete Expansion, does not appear to depend upon the nature of the query, but the use of single-term queries does considerably increase ω for the No Expansion process.

The most important single fact that can be derived from Tables 4 and 5 is that there is a substantial improvement in the relevance ratio as query expansion capability is added. The No Expansion process appears to be hopelessly inadequate for the task of document retrieval. On the other hand, the high value of ω for the Complete Expansion process dictates the need for

additional machinery to reduce the bulk of irrelevant material retrieved. Even for a message file with no multitopic messages there would, according to the histograms of Figs. 11 and 12, be a serious problem with the occasional query which produces upwards of 20 irrelevant messages.

The analysis of variance conducted on the data of Table 7 shows that some caution is needed in constructing a category-type evaluation of a retrieval system. Only after eliminating the data for three categories was it possible to accept the hypothesis that the reduced set of data were not influenced by categorization. In the case of two of the three categories, the difficulty can be traced directly to the nature of the queries associated with them. In Table 2, we find that the majority of the queries associated with Cat. 5 have a large number of factors which by Table 3, accounts for the corresponding low-relevance ratios. Similarly, we find that the preponderance of queries associated with Cat. 7 are of the single-term variety which, again by Table 3, accounts for the extraordinarily high-relevance ratios. The relevance ratio means for Cat. 3 stand out, undoubtedly because of the comparatively small number of queries used. We conclude that categorization of a message file is a valid technique for studying the relevance ratio, provided that the number of queries in each query factor group is more or less uniformly divided among the categories used.

Table 6 indicates decisively that ω cannot be measured by categorization. The erratic changes in $\overline{\omega}$ and σ_{ω} from category to category casts some doubt on the value of ω as a performance measure, although it appears to be a desirable one to use in describing the characteristics of a retrieval system. The performance measure τ is much better behaved, as indicated in Table 8, and is, perhaps, just as useful as ω in indicating the bulk of irrelevant material retrieved.

The frequency histograms in Figs. 1 through 6 exhibit much regularity in the improvement of ρ as expansion capability is added. The histograms of Figs. 7 through 12 show no such regularity in the change of the distribution of the values of ω with the addition of expansion capability. The difference between Figs. 7 and 9, or between Figs. 8 and 10 is not large, whereas there is little similarity between Figs. 9 and 11, or between Figs. 10 and 12. We conclude that the subordinate expansion has a much greater effect on ω than it does on ρ. The same conclusion can be drawn from Tables 9 and 10, which present the same data in a different way.

The importance of studying the distributions of the quantities used to measure the performance of a retrieval system can be seen in Tables 9 and 10. With respect to the Complete Expansion process, for example, we find that the probability of achieving a relevance ratio exceeding 0.9 for a given query is greater than the probability of obtaining a relevance ratio of less than 0.5 for the same query. This is a significant fact about the performance of the system.

Table 11 adequately speaks for itself, and the times for retrieval for the three processes are almost exactly as predicted by William Aldrich, who wrote the ROUT query-processing program.

The X^2-test used on the data of Table 12 reveals that the values of ρ are not significantly influenced by the people involved in the preparation of the evaluation bed, which consists of questions, rated messages, and the translations of the questions into Boolean queries.

Several different effects were studied in the sampling phase of the evaluation. The power of the Bibliography process in reducing ω was established, along with the relative weakness of the Delete process in accomplishing the same task. The translation of a question into different queries was found to have no

significant effect on the magnitude of the relevance ratio. Perhaps the most interesting aspect of the sampling study is the similarity of many of the results to those of the category testing. Relevance ratios for each process are close to the values obtained in the category testing, and, generally, are lower only because there are fewer relevant messages per question in the samples. The variations due to sampling in different ways is similar to the variations due to categorization, except that the reasons for the categorization effects are known, and the reasons for the sampling effects are not. This is, incidentally, a matter which should be intensively studied before the validity of sampling as applied to the evaluation of retrieval systems is assumed. The mean values of ω for the samples vary as erratically as they do in the case of category testing. As mentioned in the last section, the thinness of the sampling data did not make possible a detailed error analysis of the kind employed in the category testing. The weakness of the sampling technique is to some extent demonstrated in the fact that the effectiveness of the Bibliography process in reducing $\overline{\omega}$ could only be deduced from studying the ω values of Table 15, and could not be deduced from the τ values of Table 17.

The results of the last phase of the evaluation can be accepted with the most confidence, since they are based upon the rating of every message in the file against every question used. The No Expansion process was again omitted from consideration, and the Delete process, whose value was questioned in the sampling phase of the study, was further examined; (recall that an analysis of Table 14 showed that the Delete process degrades ρ to a greater extent than does the Bibliography process). The only new processes studied in this phase are the Metric Search and its extension, the Metric Search^2. The Synonym Expansion process was included so that the relative powers of the thesaurus and a purely numerical association technique could be compared.

Table 18 gives the relevance ratios for six processes, and it is immediately seen, by comparing the values for SE + BIB + METRIC and CE + BIB, that the Metric is a fair substitute for the subordinate expansion. Metric^2 provides no significant increase in capability. The Metric process was tested in the presence of a large handicap. Each of the messages in the file contains an appreciable number of words having nothing to do with the content of the message. These words are addressees, such as TAC, SAC, CONAD, etc., and they also appear in the key word dictionary. As it turns out in the study, two messages on entirely different subjects can be metrically close together simply because they have many addressees in common. The Metric process did, in spite of this handicap, boost the relevance ratio considerably. It also increased ω considerably, but there was, for a given query, much repeating of the messages retrieved. Metric searching must be studied under more reasonable conditions before statements can be made about its effect on retrieving irrelevant messages. Table 22 and Figs. 13 and 14 do, however, show that Metric Search does have a potential for increasing the capability of coordinate indexing.

A comparison of $\overline{\rho}$ for the CE + BIB process (Table 18) and the CE + DEL process (Table 19) definitely establishes the degrading effects of the Delete process, and a comparison of $\overline{\omega}$ for CE + DEL and CE + DEL + BIB in Table 19 again substantiates the superiority of the Bibliography process in cutting down ω. It may be concluded that the deletion of key words from an expanded query is not a good idea.

The discussion following Figs. 16 and 17 adequately explains the small random selection study carried out.

SECTION VIII

SUMMARY

The basic findings of the ROUT evaluation are summarized below. *

(a) Categorization is superior to sampling in the measurement of both retrieval errors and retrieval performance measures.

(b) The number of irrelevant messages retrieved is not a satisfactory performance measure, and the bulk of irrelevant material retrieved is best measured by the fraction of retrieved material which is relevant.

(c) The inherent syntactical defects of coordinate indexing produce the most significant errors.

(d) The impact of these syntactical defects is mitigated by a thesaurus of synonyms and subordinates.

(e) A thesaurus of synonym groupings provides an increase in basic performance, and is almost a necessary mechanism.

(f) The use of subordinate groupings increases performance, and these groupings can be replaced by a generalized mathematical association technique without any great loss of performance.

---

*No quantitative statements can be made about automatic indexing, since it was compared with no other kind of indexing. A qualitative appraisal of automatic indexing is that it does not limit the power of coordinate indexing retrieval, and is, hence, a satisfactory solution to the indexing problem.

(g) The power of coordinate indexing can be significantly extended by mathematical association techniques.

(h) The quality of a thesaurus has greater effects on the amount of irrelevant material retrieved than on the amount of relevant material not received.

(i) Post-retrieval filtering significantly reduces the bulk of irrelevant material retrieved.

(j) Deleting key words from query word thesaurus groupings is an inefficient and undesirable way of reducing the bulk of irrelevant material retrieved.

(k) Queries with a few multiplicative factors should be used in preference to complex queries.

(l) The complementation of query words must be done sparingly, and with caution.

(m) Increases in the performance of coordinate indexing can come about only by extending the Boolean query language.

(n) Mathematical association techniques provide a sound starting point for constructing an improved replacement for the whole coordinate indexing process.

Figure 18 summarizes, in a somewhat striking manner, the relationship of the various query processes to one another in terms of a general definition of the performance of a coordinate indexing system. A plot of ρ against τ gives a fair idea of the capability of a document retrieval system, the region of ideal performance being around the point $\rho = 1, \quad \tau = 1$.

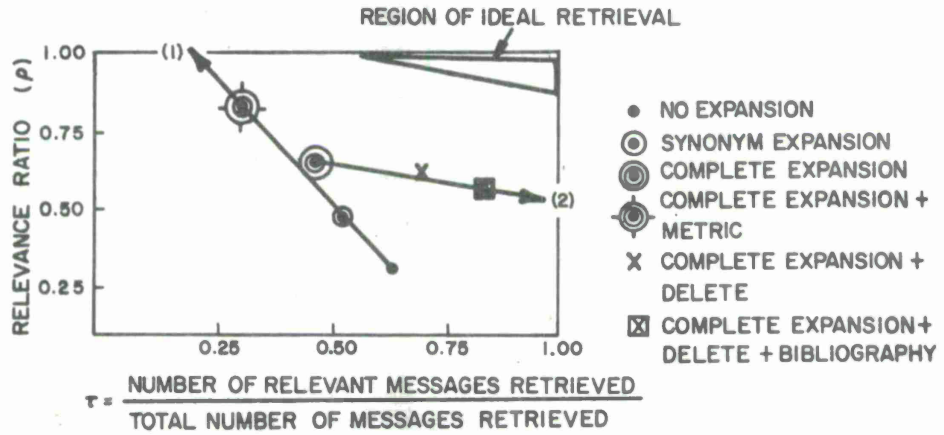


Fig. 18. Interrelationship of Query Processes

Line (1) shows the tendency to degrade τ with improvements of ρ. Line (2) shows the tendency to degrade ρ with improvement of τ. The region of ideal retrieval is approached neither by lines (1) nor (2). This is, perhaps, the most succint statement of the document retrieval problem, and Fig. 18 shows how a solution to the problem is attempted by a typical coordinate system. We thus arrive at the chief conclusion of this report:

> A generally acceptable solution to the document retrieval problem must be achieved by either vastly extending the power of the basic coordinate indexing process, or by replacing this process by an altogether different one.

J. F. Rial

BIBLIOGRAPHY

Aldrich, William J. The Query Program of ROUT 62, MITRE W-5614.

Collard, P. D. ROUT Control Program Description, MITRE W-5105.

Earnest, L. D. Intelligence Message Retrieval Experiment (Project ROTGUT), MITRE W-3779.

Famolari, E. Jr. ROUT Message Program Description, MITRE W-5179.

Fisher, R. A. Statistical Methods and Scientific Inference, Hafner Publishing Co., 1959.

Justice, C. R., Aldrich, William J., Lynch, H. W. and Radner, R. K. A Description of the ROUT System, MITRE W-5802.

Kenney, J. F. Mathematics of Statistics, Part II, D. Van Nostrand Co., 1949.

Lynch, H. W. ROUT 62 Disk Allocation, MITRE W-5565.

Mann, H. B. and Whitney, D. R. "On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other," Annals of Mathematical Statistics, Vol. 18, 1947, p. 50.

Radner, R. K. Evolution of the Dictionary and Thesaurus for the Initial Phase of ROUT 62, MITRE W-5437.

Rial, J. F. Evaluation of ROUT Coordinate Indexing System (Part II), MITRE W-5584.

Rial, J. F. A Pseudo-Metric for Document Retrieval Systems, MITRE W-4595.

Rial, J. F. Summary of ROUT Evaluation (Interim Report), MITRE W-6225.